UNIFR Did@cTIC

UNIVERSITE DE FRIBOURG, SUISSE
CENTRE DE DIDACTIQUE UNIVERSITAIRE

CONCEPTION D'UN COURS D'INFORMATIQUE SUR L'ACQUISITON DE DONNEES

Travail de fin d'étude en vue de l'obtention du Diplôme en Enseignement Supérieur et Technologie de l'Education

Sous la direction de la Prof. Bernadette Charlier Pasquier

Emmanuel DE SALIS
Filière Informatique
Groupe Analyse de données
Haute-Ecole Arc

2021

Je déclare sur mon honneur que mon travail de fin d'étude est une œuvre personnelle, composée sans concours extérieur non autorisé.

Table des matières

Résumé

Ce travail présente une partie de la création du cours Acquisition de Données. Il s'agit d'un nouveau cours de la deuxième année du Bachelor en Informatique de la Haute-Ecole Arc Ingénierie, qui sera donné pour la première fois en septembre 2021.

Concrètement, ce travail présente la conception de la table de spécifications du cours, ainsi que la création d'un travail pratique à but d'évaluation sommative. Il contient également des suggestions pour la conception du projet de module qui constitue l'évaluation sommative finale du cours.

Ce travail repose sur une première phase d'analyse des besoins auxquels ce cours doit répondre, sur la base de différents documents fournis. Cette analyse a permis d'assurer que ce cours s'inscrit bien dans le module dont il fait partie et respecte la volonté des concepteurs du module.

Glossaire

Ce travail utilise un vocabulaire spécifique propre à son contexte, reprenant celui utilisé à la Haute-Ecole Arc Ingénierie, afin de faciliter l'usage de ce travail comme guide pour le professeur donnant ce cours. Ceci garantit autant que possible que ce travail servira de base pour de futures modifications du cours.
Afin de conserver la fluidité de lecture du travail, certains points sont explicités dans le texte malgré leur présence dans le glossaire.

Concept	*Abréviation*	*Description*
Travail Pratique	TP	Evaluation sommative du cours, effectuée en partie ou totalement hors de la classe, pouvant être effectué seul·e ou en groupe selon la situation.
Travail Dirigé	TD	Evaluation formative, souvent présentée sous une forme très accompagnée (tutoriel).
Module		Blocs de plusieurs cours, régissant la jointure de leurs évaluations en une note de module, ainsi que les modalités de passage du module. Également s'y trouve la liste des prérequis du module, qui ne sont pas gérées par un cours mais bien le module dont il fait partie. Le module comprend également des objectifs généraux, qui doivent être ensuite traités par un ou plusieurs cours du module.

Introduction

Dans le cadre de sa démarche qualité, et suivant la volonté d'améliorer constamment son cursus, la Haute-Ecole Arc Ingénierie a décidé de modifier le cursus de son Bachelor en Informatique en modifiant, créant ou remplaçant certains cours. Le cours « Acquisition de données » est apparu à la suite de cette démarche, témoignant d'une volonté d'inclure un cours traitant des méthodes et problématiques relatives à l'acquisition de divers types de données informatiques.

Ce cours est une excellente opportunité d'appliquer des connaissances didactiques, car il est vierge de toute construction préalable, et permet ainsi de poser des fondations solides basées sur des concepts théoriques et pratiques. Ceci permettra notamment de garantir autant que possible sa cohérence didactique entre ses objectifs et ses méthodes d'évaluations, ainsi que fournir un cadre au professeur chargé de ce cours pour la création de futures évaluations. Une première évaluation concrète sous forme de Travail Pratique (TP), portant sur deux chapitres du cours, est également fournie dans le cadre de ce travail.

Ce travail se découpe comme suit : premièrement, une présentation du contexte permet de comprendre les objectifs exacts du travail, mais également les contraintes et libertés dont il doit tenir compte. Ensuite, la table de spécifications et son processus de création est détaillé. Un TP noté de manière sommative intervenant au milieu du semestre est ensuite proposé. Enfin, une discussion portant sur le projet de module permet de donner des suggestions sur sa conception, afin qu'elle repose sur les connaissances didactiques amenées durant ce travail.

Présentation du cours

Avant de pouvoir préparer quoi que ce soit sur ce cours, il est nécessaire de comprendre son contexte. Le module dans lequel il s'inscrit, son descriptif et celui du cours, ses objectifs, etc. sont d'autant de points à analyser afin de comprendre les contraintes de ce travail et ses axes de libertés.

Concrètement, avant de commencer ce travail, les éléments suivants ont été fourni :

1. Descriptif de module
2. Descriptif de cours
3. Description succincte des chapitres du cours (découpage par semaine)
4. Source théorique principale du cours

En plus de ceci, une séance a permis de développer les attentes des responsables de l'enseignement sur ce cours, ainsi que celles du professeur chargé du cours.

Intention du cours

A l'ère de la connexion de l'internet des objets, de plus en plus de données circulent autour de nous. Les données, ou data, sont une ressource prisée par la plupart des programmes que nous utilisons. Il est primordial pour une personne travaillant dans l'informatique de savoir d'où proviennent les données, comment a-t-on le droit de les utiliser, et dans quel but pouvons-nous les utiliser. Dans cette optique, il faut également comprendre ce que sont réellement les données, comment les acquérir, les stocker et les traiter.

Le but fondamental de ce cours est de donner des pistes de réponses à ces questions en présentant différentes problématiques et différents outils informatiques.

Descriptif de module

Le descriptif de module, disponible à l'Annexe 1 « Descriptif de module », est une version non-définitive, sujette à changement.
Sa lecture permet d'isoler les points suivants, qui servent de cadre à ce travail :

- Le cours représente 2h de cours ainsi que 2h de travail individuel par semaine
- Le module « 1242 Programmation » est un prérequis à ce module
- La langue imposée par le module est le français
- Les compétences visées par le module sont les suivantes (citées telles quelles du module) :

- *« Mettre en œuvre et exploiter des méthodes, des algorithmes et des architectures permettant le traitement, l'analyse et l'exploitation de masses de données en tenant compte des impératifs légaux, de sécurité et d'efficacité **(J)** »*
- *« Identifier les besoins, contraintes et fonctionnalités relatifs à l'acquisition, l'analyse, la gestion et le stockage d'information d'un système IT. Mettre en oeuvre ces fonctionnalités et en assurer la maintenance évolutive et corrective **(J)** »*

Le (J) fait référence à des compétences de type Jugement (analyse, synthèse, évaluation), selon la taxonomie des compétences de la Haute-Ecole Arc Ingénierie.

Descriptif de cours

Le descriptif de cours se situe à la dernière page du descriptif de module, mais pour des raisons de clarté il a été isolé dans l'Annexe 2 « Descriptif de cours ». Ce descriptif permet d'identifier les points pertinents suivants qui complètent le cadre de ce travail :

- Le langage de programmation du cours est Python et le support principal de travail sont des notebook Jupyter
- Le système d'évaluation permet bien l'usage de TP comme méthode d'évaluation

On y voit également que plusieurs objectifs sont déjà listés (cités tels quels du module) :

- *« Connaître les principes et concepts nécessaires pour une acquisition de données robuste pour faciliter les étapes d'un traitement et un stockage.*
- *Être capable de sélectionner et utiliser la typologie de données selon le traitement prévu prenant en compte aussi les impératifs légaux.*
- *Savoir utiliser de outils open source pour la récupération de données.*
- *Appliquer des approches pour l'acquisition de données depuis le web (Web Scraping).*
- *Être capable de choisir la solution d'acquisition la plus adaptée selon les besoins*
    - *Web crawling vs Web Scraping*
    - *Web scraping Vs Data mining*
    - *Data Mining Vs Process Mining*
- *Labelling et méta données être attentifs aux différentes problématiques et solutions existantes :*
    - *Définition de « label »*
    - *Data set équilibrés et déséquilibrés : problématiques et solutions*
    - *Data augmentation*
- *Droits d'utilisation des données, impératifs légaux & éthique »*

Ces objectifs serviront de base à la Table de Spécifications.

Evaluation du cours

Les étudiant·e·s seront évalués selon quatre modalités :

1. Deux TP évalués sommativement, annoncé et obligatoire, durant le semestre.
2. Projet de module en collaboration avec l'autre cours du module (Cours « Infrastructure »), intervenant dans les trois dernières semaines du semestre.
3. Un contrôle principal écrit, annoncé et obligatoire. Cette évaluation sommative fait office d'examen final et aura lieu en fin de semestre.
4. Un nombre à définir de TD durant le semestre, qui servent d'évaluation formatives.

Sources théoriques

La source théorique servant de base pour la création du cours est le livre de O'Reilly « *Data wrangling with Python: tips and tools to make your life easier* » (Kazil, J. & Jarmul, K., 2016) [3]

Ce livre traite principalement de l'aspect théorique et pratique de l'acquisition de données, mais ne parle pas d'aspect éthique et légal. A la suite d'une phase de recherche, j'ai proposé le livre de O'Reilly « Ethics of Big Data » (Davis, K. & Patterson, D., 2012) [1] pour traiter de l'aspect éthique, et ce choix a été retenu.

Pour l'aspect légal, seul le Règlement général sur la protection des données (RGPD ou GDPR en anglais) sera présenté, et le site officiel (https://gdpr-info.eu/) fera office de source.

Chapitres du cours

Le découpage en chapitre du cours a été basé sur une présentation des premiers concepts du cours et la lecture des sources théoriques mentionnées précédemment, et les points suivants ont été retenus pour ce travail (cf. Table 1) :

Semaine du semestre	**Contenu (mots-clefs)**
1	Introduction, définitions, alternatives aux base de données
2	Créations de données
3	Types de données
4	TP sur les types de données
5	Analyse de processus
6	Visualisation
7	Web Scraping (introduction)
8	Web Scraping (avancé)
9-10	TP sur le Web Scraping
11	API
12	Ethique et aspect légaux
13	Metadata et labelling
14-15-16	Projet de Module

Table 1 : Contenu du cours, par semaine

Analyse du contexte

De prime abord, deux points majeurs semblent délicats à balancer.

D'un côté, le placement du cours, c'est-à-dire le fait qu'il s'agisse d'un cours relativement petit (2h de cours par semaine) de deuxième année avec seulement le module de Programmation standard comme prérequis, et qu'il s'agit du premier cours évoquant le traitement de données, impliquent que beaucoup de concepts doivent être introduits et compris. Ceci amène beaucoup d'objectifs de type ***(C)*** *Connaissances et compréhension* et ***(A)*** *Application*, selon la taxonomie de la Haute-Ecole Arc.

D'un autre côté le descriptif de module évoquent des objectifs de type **(J)** *Jugement (analyse, synthèse, évaluation)*.

A la suite de discussions sur le sujet, il sera décidé d'inclure majoritairement des objectifs de type ***(C)*** *Connaissances et compréhension* et ***(A)*** *Application* durant les cours du semestre, et de profiter du projet de module pour évaluer les objectifs de type **(J)** *Jugement (analyse, synthèse, évaluation)*.

Table de spécifications

Cette section présente la création de la table de spécifications. Premièrement, la méthodologie de conception est détaillée afin de garantir sa validité didactique.
Deuxièmement, les différentes parties de la table de spécifications sont présentées. Les Points à Evaluer (PE) sont listés, catégorisés et filtrés, et les catégories de performance (CP) et modalités de questionnements (MQ) choisies sont présentées et expliquées.
Ceci mène à la table de spécifications finale qui a été validée par le professeur chargé du cours.

Méthodologie de conception

Une Table de Spécifications est un outil utile permettant de préciser les points à évaluer d'un cours et d'assurer la validité de celle-ci mais c'est également une source d'information pour les étudiants concernant le niveau et le type d'apprentissage attendu pour les points à évaluer mentionnés. Il est donc crucial que cette table soit créée avec une méthodologie fiable et sur des bases théoriques valides. Pour sa conception, j'ai utilisé majoritairement deux sources en plus des modules de bases de la formation didactique :

1. Le module à option B5 « Outils d'évaluation et de suivi des apprentissages en ligne » du Prof. Dr. Jean-Luc Gilles, plus particulièrement la présentation « Construction structurée des évaluations des apprentissages » (2018-2019) [2].
2. Le « Guide d'élaboration de la Table de Spécification », tiré du « Système Méthodologique d'Aide à la Réalisation de Tests (SMART) » de l'Université de Liège (2005-2006) [4].

Points à Evaluer

Catégorie de PE

Avant de définir des PE précis il convient de choisir les catégories générales dans lesquels ces PE s'inscriront. Selon le guide de l'université de Liège, il s'agit des Thèmes, que j'appellerai ici des Catégories. Ceci permet d'identifier les domaines généraux que chaque chapitre évaluera. Ainsi, après discussions avec le professeur chargé du cours, différentes catégories ont été arrêtées. Elles sont présentées dans la Table 2.

Chapitres	**Catégories**
Introduction	Définitions
	Stockage

Créations de données	Sources
Data Types	Formats
	Structuration
	Séries temporelles
	HADOOP
Acquisition de données pour l'analyse de processus	Process Mining
Bases pour la visualisation et l'analyse des données	Pandas
	Matplotlib
Intro Web Scraping	Robots.txt
	Inspection d'une page web
	Scraping
	Parsing
Web Scraping advanced solutions	Définitions
	Screen-reading
	Spider
API	API
Data and rights	Définitions
	Rights
	Ethics
Beyond data	Metadata
	Equilibrage des données

Table 2 : Présentation des catégories de PE par chapitre

Première liste de PE

A la suite de la définition des catégories, les PE peuvent être détaillés. La table 3 montre la première liste des PE retenus, par catégories.

Définitions (Introduction)	Définir "Acquisition des données"
	Définir "Big Data"
Stockage	Citer différents moyens de stocker des données
	Savoir quand utiliser une base de données (BDD)
	Citer des alternatives à une BDD
	Citer et expliquer les trois "V" (Volume, Vitesse et Variété)
Sources	Citer des sources de données

	Citer des use cases pertinents pour chaque sources de données
	Citer des avantages et inconvénients/contraintes de différentes sources de données
Formats	Reconnaître la syntaxe SQL, NoSQL, JSON, XML et CSV
	Citer les spécificités des formats de données SQL, NoSQL, JSON, XML et CSV
Structuration	Savoir les différences entre données structurées, non-structurées et semi-structurées
Time Series	Comprendre ce que sont les times series et leurs spécificités en tant que Data Type
HADOOP	Comprendre les principes de bases de HADOOP
Process Mining	Comprendre les différences entre Data Mining et Process Mining
	Savoir visualiser un processus à l'aide de ses logs
Pandas	Savoir utiliser un DataFrame et des Series pandas pour interagir avec un dataset
	Savoir importer un fichier CSV vers un DataFrame et exporter un DataFrame vers un fichier CSV
	Savoir effectuer des transformations simples sur un DataFrame pandas
Matplotlib	Afficher des données en utilisant Matplotlib
Robots.txt	Savoir à quoi sert un fichier Robots.txt et où le trouver
	Savoir lire (et écrire) un fichier Robots.txt
Inspection d'une page web	Inspecter et comprendre une hiérarchie de page html à l'aides des Developer Tools
	Identifier les API et ressources utilisées dans une page web à l'aide des Developer tools
	Interagir avec les éléments d'une page web à l'aide de JavaScript
Scraping	Utiliser les librairies urllib et urllib2 afin d'écrire des requêtes simples de scraping
	Utiliser la librairie requests pour formuler des requêtes plus complexes
Parsing	Parser une page web à l'aide de Beautiful Soup et en extraire les composants principaux
	Utiliser la librairie lxml pour lire une page web et accéder à ses composants principaux

	Utiliser XPath et regex pour pour rechercher du contenu sur une page web
Définitions (Web Scraping)	Savoir quels sont les principaux types de Scraper (page reading, screen-reading et spider) et quels sont leurs cas d'application
Screen-reading	Faire du screen-reading avec Selenium
	Faire du screen-reading avec Ghost.py
Spider	Connaître les différents types de spider et leur cas d'application
	Construire un spider basique avec Scrapy
	Implémenter de la gestion d'erreur dans son spider
API	Connaître ce qu'est une API
	Connaître les différences entre une API REST et Streaming
	Savoir interagir avec une API pour obtenir des données
Définitions (Data and rights)	Connaître les différences entre Open Data et Closed Data
	Connaître les différentes licences
Rights	Savoir ce qu'est la GDPR et ses principes majeurs
Ethics	Connaître les avantages de l'enquête/analyse éthique
	Connaître les 4 éléments de l'éthique du Big-Data
	Connaître les points de décisions éthiques
Metadata	Comprendre les différences entre data et metadata
Equilibrage des données	Savoir ce qu'est un dataset balancé et non-balancé
	Connaître les risques liés à un dataset non-balancé
	Citer des pistes pour balancer un dataset
	Connaître des techniques de base de Data Augmentation (Random noise, SMOTE, etc.)

Table 3 : Liste des PE, groupés par catégories

Tri et liste finale retenue

À la suite de plusieurs revues et discussions, la liste des PE a été raffinée et améliorée. Afin d'éviter de la redondance dans le rapport, la liste finale est visible directement dans la section « Table de Spécifications finale », plus loin dans le rapport.

Catégorie de Performance

Concernant les CP, les suivantes ont été retenues :

- **Connaissance**
- **Compréhension**
- **Application**

- **Analyse**
- **Synthèse**
- **Evaluation**

Ce choix est motivé par deux facteurs : ils sont préconisés par le guide de l'université de Liège, et rejoignent les trois types d'objectifs de la Haute-Ecole Arc : ***(C)*** *Connaissances et compréhension,* ***(A)*** *Application* et **(J)** *Jugement (analyse, synthèse, évaluation).* La correspondance se lit donc comme suit :

- **Connaissance (C)**
- **Compréhension (C)**
- **Application (A)**
- **Analyse (J)**
- **Synthèse (J)**
- **Evaluation (J)**

Modalité de Questionnement

Dans la version actuelle de la table de spécifications, les modalités de questionnement (MQ) sont définies par CP. Cela signifie qu'une CP aura toujours la même MQ. Cela pourra être amené à changer dans une version future. Les MQ attribuées à chaque CP sont présentées dans la table 4

Connaissance	Questions à choix multiples (QCM)
Compréhension	Questions à choix multiples (QCM)
Application	Questions ouvertes à réponses courtes (QROC)
Analyse	Questions ouvertes à réponses longues (QROL)
Synthèse	Questions ouvertes à réponses longues (QROL)
Evaluation	Questions ouvertes à réponses longues (QROL)

Table 4 : Présentation des MQ choisies par CP

Table de spécifications finale

Ces différentes étapes ont permis d'arriver à la Table de Spécifications, présentée dans la table 5 ci-dessous.

Points à évaluer		**Catégories de performances**			
		Connaissance	**Compréhension**	**Application**	**Analyse**
Introduction					
Définitions	Définir "Acquisition des données"	x			

	Définir "Big Data"	x			
Stockage	Citer différents moyens de stocker des données	x			
	Analyser le besoin d'utiliser une base de données (BDD)				x
	Citer des alternatives à une BDD	x			
	Citer et expliquer les trois "V" (Volume, Vitesse et Variété)		x		
Créations de données					
Sources	Citer des sources de données	x			
	Développer des use cases pertinents pour chaque sources de données		x		
	Citer des avantages et inconvénients/contraintes de différentes sources de données	x			
Data Types					
Formats	Reconnaître la syntaxe SQL, NoSQL, JSON, XML et CSV	x			
	Détailler les spécificités des formats de données SQL, NoSQL, JSON, XML et CSV		x		
Structuration	Savoir les différences entre données structurées, non-structurées et semi-structurées	x			
Time Series	Comprendre ce que sont les times series et savoir les traiter			x	
HADOOP	Comprendre les principes de bases de HADOOP et les appliquer			x	
Acquisition de données pour l'analyse de processus					
Process Mining	Comprendre les différences entre Data Mining et Process Mining	x			
	Visualiser un processus à l'aide de ses logs			x	
Bases pour la visualisation et l'analyse des données					
Pandas	Utiliser un DataFrame et des Series pandas pour interagir avec un dataset			x	
	Importer un fichier CSV vers un DataFrame et exporter un DataFrame vers un fichier CSV			x	
	Effectuer des transformations simples sur un DataFrame pandas			x	
Matplotlib	Afficher des données en utilisant Matplotlib			x	

Introduction Web Scraping					
Robots.txt	Savoir à quoi sert un fichier Robots.txt et où le trouver	x			
	Lire (et écrire) un fichier Robots.txt		x		
Inspection d'une page web	Inspecter et comprendre une hiérarchie de page html à l'aides des Developer Tools			x	
	Identifier les API et ressources utilisées dans une page web à l'aide des Developer tools				x
	Interagir avec les éléments d'une page à l'aide de JavaScript			x	
Scraping	Utiliser les librairies urllib et urllib2 afin d'écrire des requêtes simples de scraping			x	
	Utiliser la librairie requests pour formuler des requêtes plus complexes			x	
Parsing	Parser une page web à l'aide de Beautiful Soup et en extraire les composants principaux			x	
	Utiliser la librairie lxml pour lire une page web et accéder à ses composants principaux			x	
	Utiliser XPath et regex pour pour rechercher du contenu sur une page web			x	
Web Scraping advanced solutions					
Définitions	Savoir quels sont les principaux types de Scraper (page reading, screen-reading et spider) et quels sont leurs cas d'application	x			
Screen-reading	Faire du screen-reading avec Selenium			x	
	Faire du screen-reading avec Ghost.py			x	
Spider	Connaître les différents types de spider et leur cas d'application	x			
	Construire un spider basique avec Scrapy			x	
	Implémenter de la gestion d'erreur dans son spider			x	
API					
API	Connaître ce qu'est une API	x			
	Détailler les différences entre une API REST et Streaming		x		

	Interagir avec une API pour obtenir des données			x	
Data and rights					
Définitions	Connaître les différences entre Open Data et Closed Data	x			
	Connaître les différentes licences	x			
Rights	Savoir ce qu'est la GDPR et ses principes majeurs	x			
Ethics	Connaître les avantages de l'enquête/analyse éthique	x			
	Connaître les 4 éléments de l'éthique du Big-Data	x			
	Connaître les points de décisions éthiques	x			
Beyond data					
Metadata	Comprendre les différences entre data et metadata	x			
Equilibrage des données	Savoir ce qu'est un dataset balancé et non-balancé	x			
	Connaître les risques liés à un dataset non-balancé		x		
	Citer des pistes pour balancer un dataset		x		
	Appliquer des techniques de base de Data Augmentation (Random noise, SMOTE, etc.)			x	

Table 5 : Table de Spécifications, première version.

Cette version contient une modification par rapport aux choix établis précédemment. En effet, aucun PE n'a de CP de type Synthèse ou Evaluation, ces deux CP ont donc été supprimé de cette version de la Table de Spécifications afin de l'alléger. Néanmoins, il n'est pas exclu que certains PE changent de CP dans le futur, ou que de nouveaux PE apparaissent.

TP 1 - Web Scraping

Sélection de la table de spécification relative au TP

La partie évaluée par ce TP est principalement la section de la Table de Spécifications du chapitre « Introduction Web Scraping » et « Bases pour la visualisation et l'analyse des données » (cf. Table 5) mais également certains points des chapitres précédents. Les PE retenus sont :

- Inspecter et comprendre une hiérarchie de page html à l'aides des Developer Tools
- Utiliser la librairie requests pour formuler des requêtes plus complexes
- Parser une page web à l'aide de Beautiful Soup et en extraire les composants principaux
- Utiliser un DataFrame et des Series pandas pour interagir avec un dataset
- Importer un fichier CSV vers un DataFrame et exporter un DataFrame vers un fichier CSV
- Effectuer des transformations simples sur un DataFrame pandas
- Afficher des données en utilisant Matplotlib

Réalisation du TP

Les PE sélectionnés ayant uniquement des CP de type Application, ce TP sera sous forme de programme à écrire, à l'aide des outils préconisés par le descriptif de module. La donnée TP a été exporté pour pouvoir être ajouté au rapport, et est visible dans la Figure 1 ci-dessous. La lecture des notes de cours du module Module B. « Evaluation des apprentissages » a été utile pour guider la rédaction de ce TP.

Le corrigé est disponible en annexe (Annexe 3 « *Corrigé TP Web Scraping* »). Certains points peuvent sembler différents entre le corrigé et la donnée, et ceci est dû au fait que le corrigé n'est pas destiné à être donné tel quel aux étudiants, mais plutôt détaillé en classe une fois le travail rendu.

Acquisition de Données

TP Web Scraping

Objectif: *Créer une base de données composée des faits divers "Le saviez-vous?" affichés en page d'accueil de Wikipédia, de 2017 à 2021*

- Setup -

Les lignes expliquent servent de tutoriel afin de réaliser pas-à-pas la mise en place nécessaire pour ce TP
Les commandes sont données pour Windows et utilisent pip, mais si vous voulez utiliser d'autres OS ou outils (conda, etc.), libres à vous, tant que vous pouvez faire le TP.

1. **Créez un environnement virtuel Python pour ce TP, à l'emplacement de votre choix**

python -m venv scraping-venv

1. **Activez-le**

scraping-venv\Scripts\activate.bat

1. **Installez-y les packages requis pour le TP, ainsi que ipykernel**

pip install requests

pip install bs4

...

1. **Transformez le venv en kernel Jupyter**

ipython kernel install --user --name=scraping-kernel

1. **Voilà, vous devriez pouvoir utiliser ce kernel dans votre Jupyter-lab**

- Imports -

Voici la liste des imports dont vous devriez avoir besoin, vous pouvez en ajouter ou enlever si vous le jugez nécessaire

In [2]:
```
import requests
from bs4 import BeautifulSoup
import pandas as pd
import dateparser
import matplotlib.pyplot as plt
```

- Scraping the first page -

- Tout d'abord, allez sur https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Le_saviez-vous_%3F/Archives (https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Le_saviez-vous_%3F/Archives) pour vous familiariser avec la page.
- Utilisez requests pour scraper la page.
- Utilisez BeautifulSoup pour parser le contenu en un objet BeautifulSoup

Utilisez autant de cellules que nécessaires, et n'oubliez pas de commenter votre code.

- Identifying relevant page part and first quote -

- Utilisez l'inspecteur de votre navigateur web pour trouver où se situe le contenu de la page, et les quotes.
- Utilisez les propriétés de BeautifulSoup pour isoler la partie pertinente
- Isoler la première quote et affichez-la

Utilisez autant de cellules que nécessaires, et n'oubliez pas de commenter votre code.

- Creating pandas DataFrame -

- Créez un DataFrame pandas afin de récolter les différentes quotes que vous allez parser.
- Utilisez les colonnes fournies
- Ajoutez la première quote précédemment extraite comme première entrée de votre DataFrame

Utilisez autant de cellules que nécessaires, et n'oubliez pas de commenter votre code.

In [1]:
```
#Noms des colonnes du dataframe pandas
columns=['year', 'date', 'quote']
```

- Parsing 2021 -

- L'objectif de cette section est de parser l'année 2021 complète et de l'ajouter au DataFrame
- Les fonction find, rfind, etc. des strings Python peuvent vous être utiles pour identifier les parties pertinentes des quotes
- dateparser vous permet de transformer une date sous forme de texte en une date à un format plus standard

Utilisez autant de cellules que nécessaires, et n'oubliez pas de commenter votre code.

- Parsing 2017 to 2020 -

- L'objectif de cette section est de parser les années 2017 à 2020 et de les ajouter au DataFrame
- Les fonction find, rfind, etc. des strings Python peuvent vous être utiles pour identifier les parties pertinentes des quotes
- dateparser vous permet de transformer une date sous forme de texte en une date à un format plus standard

Utilisez autant de cellules que nécessaires, et n'oubliez pas de commenter votre code.

- df Transformations -

- Supprimez la colonne 'year' du DataFrame
- Importez de la colonne 'date' des informations permettant de générer une colonne 'year' et 'month'

Utilisez autant de cellules que nécessaires, et n'oubliez pas de commenter votre code.

- Exporting DataFrame to csv -

- Exporter votre DataFrame avec le nom donné
- Utilisez ';' comme séparateur et n'exporter pas l'index

Utilisez autant de cellules que nécessaires, et n'oubliez pas de commenter votre code.

In [2]:
```
nom_csv = 'Wikiquotes.csv'
```

- Statistics -

- Utilisez les fonctions ou propriétés de votre choix afin de répondre aux questions suivantes

In [4]:
```
# 0) Y a-t-il des lignes pour lesquelles certains champs sont vides? Si oui, combien?
```

```
In [5]: # 1) Quelle année contient le plus de quotes, et combien?

In [6]: # 2) Quelle année contient le moins de quotes, et combien?

In [7]: # 3) Quelle est la quote la plus longue? (En terme de nombre de cara
        ctères)

In [8]: # 4) Quelle est la quote la plus courte? (En terme de nombre de cara
        ctères)

In [9]: # 5) Combien de quotes contiennent 'Victor Hugo'?

In [10]: # 6) J'ai entendu dire qu'il y avait une quote qui parlait de pokémo
         n, laquelle est-ce? (Affichez la quote complète)
```

- Visualisation -

- Affichez un graphe de votre choix pour visualiser les données, que vous jugez pertinent (par exemple un barplot, scatterplot, ou autre)

Figure 1 : Donnée du TP Web Scraping

Projet de Module

Suggestion(s) de projet

Le projet de module doit reprendre une évaluation du cours d'Acquisition de Données, mais également de l'autre cours du module, soit le cours « Infrastructure ». Malheureusement le temps disponible pour la réalisation de ce travail n'a pas permis de fournir un prototype de projet de module sous forme de programme, mais plusieurs idées ont été arrêtées et peuvent servir de piste pour la conception de ce projet de fin de semestre :

1. Evaluer la partie avancée du Web Scraping avec la création d'un « spider », puis sauvegarde des données obtenues sous un format suggéré par le cours « Infrastructure ». Le sujet du scraping sera choisi individuellement par les étudiants, avec validation par le professeur chargé du cours.
2. Apprentissage de l'API Twitter pour faire des requêtes permettant de sauvegarder une grande quantité de tweets sur un thème donné, puis sauvegarder des données obtenues sous un format suggéré par le cours « Infrastructure ». Enfin, une analyse des tendances du thème sera à produire.

Il serait également intéressant de profiter du projet de module pour ajouter des éléments de type Jugement, tel que suggéré par le descriptif de module. En l'état, il n'y a pas de PE qui bénéficient de CP de type Jugement, mais certains points s'y prêtent particulièrement, tels que les dimensions de droits et d'éthique abordées dans le chapitre « Data and rights ». Une solution possible serait d'ajouter une évaluation critique du projet de module d'un point de vue éthique, à rendre avec le projet. Dans ce cas, il serait nécessaire de modifier la table de spécifications pour revoir les PE en question, et leur attribuer une CP adéquate (analyse, synthèse ou évaluation).

Conclusion

Ce travail détaille la conception de la table de spécifications d'un nouveau cours encore jamais enseigné, ainsi qu'une évaluation sommative. Des idées pour un projet de fin de semestre sont également proposées.

La table de spécifications bien qu'encore amenée à évoluer, est déjà assez complète pour aider à la création du support de cours et de son contenu. Je ne peux garantir qu'elle sera fournie aux étudiants du cours, mais elle a été créé dans cette optique, et la prochaine étape logique serait d'évaluer son impact perçu par les étudiants sur leur apprentissage à la fin du semestre.

D'un point de vue personnel, ce travail m'aura parmi d'appliquer plusieurs concepts vus durant l'ensemble de ma formation didactique, et je suis particulièrement reconnaissant à la Haute-Ecole Arc de m'avoir permis de m'y atteler sur un cours entièrement nouveau. Les échanges avec le professeur chargé du cours ont été instructifs et m'ont permis de comprendre les points nécessaires pour transformer des concepts théoriques à une application concrète dans un contexte professionnel. Ceci, bien que transparaissant moins dans ce travail, constitue néanmoins un apprentissage fort dont je suis très reconnaissant.

Bibliographie

[1] Davis, K., & Patterson, D. (2012). *Ethics of big data:*. Sebastopol, CA: OReilly.

[2] Gilles, J-L., (2018) *Construction structurée des évaluations des apprentissages* [PDF Slides] https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxvZXNhZWx8Z3g6MWI0MmI4MmFhNTJjYWQ2Yw

[3] Kazil, J., & Jarmul, K. (2016). *Data wrangling with Python: Tips and tools to make your life easier*. Sebastopol, CA: OReilly.

[4] Système Méthodologique d'Aide à la Réalisation de Tests (SMART) (2005) *Guide d'élaboration de la Table de Spécification*. Université de Liège

Annexe

Annexe 1. Descriptif de module

Descriptif de module

Domaine : Ingénierie et Architecture

Filière : Informatique et système de communication (ISC)

Orientation : Ingénierie des données (ID)

Axe de Formation : Bases des TIC

Intitulé de module **Ingénierie des données SA**

Code : 2271

Type de formation : ☒ Bachelor ☐ Master ☐ MAS ☐ DAS ☐ CAS ☐ Autres :

Langue principale d'enseignement : ☒ Français ☐ Anglais ☐ Allemand

Organisation

Crédits ECTS : **7**

Unités d'enseignement :

			1		2		3	
		Période pédagogique (semestre)	1	2	3	4	5	6
N°	**Type**	**Désignation**						
2271.1	CT	Introduction aux données II (cours bloc) HES d'été						
2271.2	CT	Infrastructure			4			
2271.3	CT	Acquisition des données			2			
	Examen							
		Total			6			

Indication en périodes d'enseignement hebdomadaires (45 min.)

CT – Cours théorique ; TP - Travail pratique ; PR – Projet

Volume de travail :

	heures
Enseignement	68
Travail personnel	142
Travail total	210

1. Prérequis

☒ Avoir validé le module : 1242 Programmation
☐ Avoir suivi le module
☐ Pas de prérequis
☐ Autres :

2. Compétences visées

Compétences visées par le module	*Les objectifs d'apprentissage de ce module sont classés selon les trois degrés croissants de difficulté:* ***(C)*** *Connaissances et compréhension* ***(A)*** *Application,* ***(J)*** *Jugement (analyse, synthèse, évaluation).* **A l'issue du module, l'étudiant est capable de :** - Mettre en oeuvre et exploiter des méthodes, des algorithmes et des architectures permettant le traitement, l'analyse et l'exploitation de masses de données en tenant compte des impératifs légaux, de sécurité et d'efficacité ***(J)*** - Identifier les besoins, contraintes et fonctionnalités relatifs à l'acquisition, l'analyse, la gestion et le stockage d'information d'un système IT. Mettre en oeuvre ces fonctionnalités et en assurer la maintenance évolutive et corrective ***(J)***

Modalités d'évaluation et de validation

Evaluation des apprentissages

- Evaluations des différentes Unités d'Enseignement (UE)

Note finale du module :

$$M = \frac{m_{IDOII} + m_{INFR} + m_{ADO}}{3}$$

m_{IDOII} = moyenne des notes d'Introduction aux données II
m_{INFR} = moyenne des notes d'Infrastructure
m_{ADO} = moyenne des notes d'Acquisition des données

Toutes les notes et moyennes sont précisées au dixième de point.

Conditions de réussite :

Note finale du module	$M \geq 4.0$ (arrondie au demi-point)
Moyenne de chacune des UE	$m_i \geq 3.0$ (arrondies au dixième de point)

La note finale du module, calculée au dixième de point, permet d'établir la note ECTS.

Modalités de remédiation

☐ Remédiation possible
☒ Pas de remédiation
☐ Autre (précisez) : …

Modalités de répétition

L'étudiant qui répète un module ne refait pas les unités d'enseignement du module dont la moyenne m_i est égale ou supérieure à 5.0 arrondi au ½ point. Sur demande l'étudiant peut refaire une unité d'enseignement à laquelle il n'est pas astreint.

Contenu et formes d'enseignement

Unité d'enseignement	Introduction aux données II (cours bloc) HES d'été
Identifiant	2271.1
Méthode d'enseignement	Cours, exercices et mini-projet de groupe.
Objectifs spécifiques	- Comprendre le cycle de vie d'un projet d'ingénierie de données et les tâches liées aux phases d'analyse, conception, développement, test - Conception : diagrammes et documents, argumentation des choix - Comprendre les différences entre les principaux patterns d'architecture logicielle - Approfondir les bonnes pratiques d'utilisation de versionning et de forge Comprendre les avantages d'une bonne conception pour la réalisation d'un projet
Modalités d'évaluation	La note d'évaluation de ce cours bloc portera essentiellement sur les tâches de gestion de projet et de conception pour la réalisation du mini-projet. Tous les résultats et délivrables se trouveront sur la forge du projet
Description du contenu (mots clés)	Gestion de projet, Analyse, Conception, Documentation, Planification, Versionning, Forge, Wiki, Diagrammes, UML
Supports de cours	Au choix des enseignants
Outils utilisés	Git, gitlab, markdown, C++, Qt
Bibliographie	
Particularité d'organisation	Cours bloc de 11 ou 12 jours

Unité d'enseignement	Infrastructure
Identifiant	2271.2
Méthode d'enseignement	Cours théoriques et exercices pratiques en classe
Objectifs spécifiques	• Comprendre les concepts de base de l'infrastructure des systèmes de calcul en nuages. • Être en mesure de déployer des applications d'analyse de données à grande échelle. • Savoir utiliser les infrastructures de calcul en large échelle pour développer des applications d'analyse de données textuels, des graphes et des données relationnelles.
Modalités d'évaluation	• Un contrôle principal écrit, annoncé et obligatoire. • Mini-projet de module avec l'unité d'enseignement Acquisition de données
Description du contenu (mots clés)	Introduction à l'infrastructure de Big Data pour l'analyse de données. Conception d'algorithmes et la réflexion à l'échelle. Techniques de support de calcul textuel, des graphes et des données relationnelles. Abstractions MapReduce, Spark et streaming. Infrastructure de déploiement et d'exécution basée sur les machines virtuelles et les containers. Infrastructures de stockage à large échelle (bases de données distribuées, noSQL).
Supports de cours	• Au choix de l'enseignant.
Outils utilisés	• Hadoop, Spark • Java et Python • Services de gestion de cloud
Bibliographie	Tom White. Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale. O'Reilly Media; 4th edition (April 21, 2015). Bill Chambers, Matei Zaharia. Spark: The Definitive Guide: Big Data Processing Made Simple. O'Reilly Media (March 20, 2018).
Particularité d'organisation	Aucune.

Unité d'enseignement	Acquisition des données
Identifiant	2271.3
Méthode d'enseignement	Cours théoriques et exercices pratiques en classe

Objectifs spécifiques	• Connaître les principes et concepts nécessaires pour une acquisition de données robuste pour faciliter les étapes d'un traitement et un stockage.
	• Être capable de sélectionner et utiliser la typologie de données selon le traitement prévu prenant en compte aussi les impératifs légaux.
	• Savoir utiliser de outils open source pour la récupération de données.
	• Appliquer des approches pour l'acquisition de données depuis le web (Web *Scraping*).
	• Être capable de choisir la solution d'acquisition la plus adaptée selon les besoins
	  ○ Web crawling vs Web Scraping
	  ○ Web scraping Vs Data mining
	  ○ Data Mining Vs Process Mining
	• *Labelling* et méta données être attentifs aux différentes problématiques et solutions existantes :
	  ○ Définition de « label »
	  ○ Data set équilibrés et déséquilibrés : problématiques et solutions
	  ○ *Data augmentation*
	• Droits d'utilisation des données, impératifs légaux & éthique

Modalités d'évaluation	• Un contrôle principal écrit, annoncé et obligatoire. • Possibilité de travaux pratiques évalués (mini-projets) • Mini-projet de module avec l'unité d'enseignement Infrastructure
Description du contenu (mots clés)	*Big data* vs *small data* ; acquisition vs création de données (et augmentation des données) ; données structurées, non-structurées et semi-structurées (y.c., JSON et XML); *web scraping* ; metadata & labelling ; droits d'utilisation des données, impératifs légaux & éthique ; bases d'analyse et visualisations de données
Supports de cours	• Slides (Jupyter) Notebooks
Outils utilisés	• Python (scripts & Jupyter Notebooks) Bibliothèques existantes
Bibliographie	Kazil, J., & Jarmul, K. (2016). *Data wrangling with Python: tips and tools to make your life easier*. "O'Reilly Media, Inc."
Particularité d'organisation	Aucune.

Annexe 2 « Descriptif de cours »

Unité d'enseignement	**Acquisition des données**
Identifiant	2271.3
Méthode d'enseignement	Cours théoriques et exercices pratiques en classe

Objectifs spécifiques	• Connaître les principes et concepts nécessaires pour une acquisition de données robuste pour faciliter les étapes d'un traitement et un stockage.
	• Être capable de sélectionner et utiliser la typologie de données selon le traitement prévu prenant en compte aussi les impératifs légaux.
	• Savoir utiliser de outils open source pour la récupération de données.
	• Appliquer des approches pour l'acquisition de données depuis le web (Web *Scraping*).
	• Être capable de choisir la solution d'acquisition la plus adaptée selon les besoins
	  ○ Web crawling vs Web Scraping
	  ○ Web scraping Vs Data mining
	  ○ Data Mining Vs Process Mining
	• *Labelling* et méta données être attentifs aux différentes problématiques et solutions existantes :
	  ○ Définition de « label »
	  ○ Data set équilibrés et déséquilibrés : problématiques et solutions
	  ○ *Data augmentation*
	• Droits d'utilisation des données, impératifs légaux & éthique

Modalités d'évaluation	• Un contrôle principal écrit, annoncé et obligatoire. • Possibilité de travaux pratiques évalués (mini-projets) • Mini-projet de module avec l'unité d'enseignement Infrastructure
Description du contenu (mots clés)	*Big data* vs *small data* ; acquisition vs création de données (et augmentation des données) ; données structurées, non-structurées et semi-structurées (y.c., JSON et XML); *web scraping* ; metadata & labelling ; droits d'utilisation des données, impératifs légaux & éthique ; bases d'analyse et visualisations de données
Supports de cours	• Slides (Jupyter) Notebooks
Outils utilisés	• Python (scripts & Jupyter Notebooks) Bibliothèques existantes
Bibliographie	Kazil, J., & Jarmul, K. (2016). *Data wrangling with Python: tips and tools to make your life easier*. "O'Reilly Media, Inc."
Particularité d'organisation	Aucune.

Annexe 3 « Corrigé TP Web Scraping »

Acquisition de Données

TP Web Scraping

Objectif: *Créer une base de données composée des faits divers "Le saviez-vous?" affichés en page d'accueil de Wikipédia, de 2017 à 2021*

- Imports -

In [31]:
```
import requests
from bs4 import BeautifulSoup
import pandas as pd
import dateparser
import matplotlib.pyplot as plt
```

- Scraping the first page -

In [71]:
```
r = requests.get('https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Le_saviez-vous_%3F/Archives')
print(type(r))
```

```
<class 'requests.models.Response'>
```

In [72]:
```
print(r)
```

```
<Response [200]>
```

In [73]:
```
content = r.content
print(type(content))
```

```
<class 'bytes'>
```

In [ ]:
```
print(content)
```

In [74]:
```
soup = BeautifulSoup(content)
print(type(soup))
```

```
<class 'bs4.BeautifulSoup'>
```

In [ ]:
```
print(soup)
```

- Identifying relevant page part and first quote -

In [75]:
```
#Done using the web inspector
subsoup = soup.find('div', class_='mw-parser-output')
```

In [76]:
```
type(subsoup)
```

Out[76]:
```
bs4.element.Tag
```

In [77]:
```
#print(subsoup)
```

In [78]:
```
first_entry = subsoup.find_all('p')[-1].text
first_entry
```

Out[78]:
```
'Victor Hugo a empêché pendant six ans la représentation en France
de Rigoletto (affiche du haut) de Giuseppe Verdi, estimant qu'il
s'agissait d'un plagiat d'une de ses œuvres (affiche du bas).\n'
```

In [79]:
```
subsoup.find_all('li')[1].text
```

Out[79]:
```
"Au pays des kiwis, la France se nomme Wīwī.\nprésente 2 jours en
page d'accueil du 01 janvier 2021 à 00:00:22 au 03 janvier 2021 à
00:00:13. Discussion de la proposition."
```

- Creating pandas DataFrame -

In [80]:
```
df = pd.DataFrame(columns=['year', 'date', 'quote'])
```

In [81]:
```
df = df.append({
        'year': 2021,
        'date': dateparser.parse('2021-01-01'),
        'quote': first_entry.strip()
    }, ignore_index = True)
```

In [82]:
```
df
```

Out[82]:

	year	date	quote
0	2021	2021-01-01	Victor Hugo a empêché pendant six ans la repré...

- Parsing 2021 -

In [83]:
```
year = 2021
first = True
for li in subsoup.find_all('li'):
    #print('\n')
    #print(li.text)
    if first:
        first = False
        continue
    text = li.text
    quote_end_index = text.rfind('présente')
    quote = text[:quote_end_index]
    date_start_index = text.find('du', quote_end_index) + 3
    date_end_index = text.find('2021', quote_end_index) + 4
    date = dateparser.parse(text[date_start_index:date_end_index])

    df = df.append({
        'year': year,
        'date': date,
        'quote': quote.strip()
    }, ignore_index = True)
```

In [84]:
```
df
```

Out[84]:

	year	date	quote
0	2021	2021-01-01	Victor Hugo a empêché pendant six ans la repré...
1	2021	2021-01-01	Au pays des kiwis, la France se nomme Wīwī.
2	2021	2021-01-01	Réalisés le même jour, La Rue Mosnier aux drap...
3	2021	2021-01-02	L'astéroïde (24601) Valjean est ainsi nommé en...
4	2021	2021-01-03	D'abord hôtel particulier, désormais résidence...
...	...	...	...
563	2021	2021-07-16	Aucun des prix du Festival de Cannes ne fut dé...
564	2021	2021-07-17	Créée en 1955 par une femme, la Palme d'or (ph...
565	2021	2021-07-17	En 1945, Luigi Durand de la Penne a reçu la pl...
566	2021	2021-07-18	En 1904, Henri Cornet (photo) a été désigné va...
567	2021	2021-07-18	Passé au français, le breton Morbihan perd son...

568 rows × 3 columns

- Parsing 2017 to 2020 -

In [85]:
```
year = range(2017, 2021)
for y in year:
    soup = BeautifulSoup(requests.get('https://fr.wikipedia.org/wiki
/Wikip%C3%A9dia:Le_saviez-vous_%3F/Archives/' + str(y)).content)
    subsoup = soup.find('div', class_='mw-parser-output')
    for li in subsoup.find_all('li'):
        #print('\n')
        #print(li.text)
        text = li.text
        quote_end_index = text.rfind('présente')
        quote = text[:quote_end_index]
        date_start_index = text.find('du', quote_end_index) + 3
        date_end_index = text.find(str(y), quote_end_index) + 4
        date = dateparser.parse(text[date_start_index:date_end_inde
x])

        df = df.append({
            'year': y,
            'date': date,
            'quote': quote.strip()
        }, ignore_index = True)
```

In [86]:
```
df.tail()
```

Out[86]:

	year	date	quote
4691	2020	2020-12-30	L'année de la sortie d'À bout de souffle, film...
4692	2020	2020-12-30	Ces funérailles musicales (son) ne sont pas un...
4693	2020	2020-12-31	Dans le calendrier cosmique de Carl Sagan, l'H...
4694	2020	2020-12-31	L'augmentation de CO2 dans l'atmosphère consti...
4695	2020	2020-12-31	S'il ne porte pas de millésime sur la bouteill...

- Exporting DataFrame to csv -

In [87]:
```
df.to_csv('Wikiquotes.csv', sep=';', index=False)
```

In [3]:
```
df = pd.read_csv('Wikiquotes.csv', sep=';')
df.head()
```

Out[3]:

	year	date	quote
0	2021	2021-01-01	Victor Hugo a empêché pendant six ans la repré...
1	2021	2021-01-01	Au pays des kiwis, la France se nomme Wīwī.
2	2021	2021-01-01	Réalisés le même jour, La Rue Mosnier aux drap...
3	2021	2021-01-02	L'astéroïde (24601) Valjean est ainsi nommé en...
4	2021	2021-01-03	D'abord hôtel particulier, désormais résidence...

- df Transformations -

In [4]:
```
df = df.drop(columns=['year'])
```

In [5]:
```
df['date'] = pd.to_datetime(df['date'])
df['year'], df['month'] = df['date'].dt.year, df['date'].dt.month
df.head()
```

Out[5]:

	date	quote	year	month
0	2021-01-01	Victor Hugo a empêché pendant six ans la repré...	2021.0	1.0
1	2021-01-01	Au pays des kiwis, la France se nomme Wīwī.	2021.0	1.0
2	2021-01-01	Réalisés le même jour, La Rue Mosnier aux drap...	2021.0	1.0
3	2021-01-02	L'astéroïde (24601) Valjean est ainsi nommé en...	2021.0	1.0
4	2021-01-03	D'abord hôtel particulier, désormais résidence...	2021.0	1.0

- Statistics -

In [6]:
```
df.describe()
#mmmh, not very useful
```

Out[6]:

	year	month
count	4692.000000	4692.000000
mean	2018.872975	6.439898
std	1.311144	3.369204
min	2017.000000	1.000000
25%	2018.000000	4.000000
50%	2019.000000	6.000000
75%	2020.000000	9.000000
max	2021.000000	12.000000

In [7]:
```
#First statistical analysis/checks:
#Are there rows with missing values in any columns? How many?
df[df.isnull().any(1)]
```

Out[7]:

	date	quote	year	month
568	NaT	Créée en 1829, la Revue des deux Mondes (repro...	NaN	NaN
569	NaT	En Suède, les règles d'usage du tutoiement et ...	NaN	NaN
570	NaT	En Allemagne, la Citroën 2 CV (photo) est surn...	NaN	NaN
571	NaT	En décembre 1916, ayant découvert que la densi...	NaN	NaN

In [16]:
```
#Use pandas to answer those questions:
#1) What year had the most quotes and how much?
df.year.value_counts().head(1)
```

Out[16]:
```
2018.0    1250
Name: year, dtype: int64
```

In [9]:
```
#2) What year had the least quotes and how much?
df.year.value_counts().tail(1)
```

Out[9]:
```
2021.0    568
Name: year, dtype: int64
```

In [34]:
```
#3) What is the longest quote?
max(df.quote, key=len)
```

Out[34]:
```
'L'escargot chante\xa0; respire à côté de son anus\xa0; sent sous
ses yeux et voit (mal) avec ses tentacules rétractiles supérieurs\
xa0; touche avec ses tentacules inférieurs\xa0; mange avec sa lang
ue garnie de dents située sous son pied\xa0; peut se déplacer (len
tement) sur des lames de rasoir sans se blesser grâce à son mucus\
xa0; et porte un vagin et un pénis utilisés simultanément, tout en
mutilant son partenaire avec un dard d'amour.'
```

In [11]:
```
#4) What is the shortest quote?
min(df.quote, key=len)
```

Out[11]:
```
'Film en est un.'
```

In [12]:
```
#5) How many quotes contains 'Victor Hugo'?
len(df[df.quote.str.contains('Victor Hugo')])
```

Out[12]:
```
15
```

In [13]:
```
#6) I heard there is a quote talking about pokémon, what is it? (Pri
nt whole quote)
df[df.quote.str.contains('pokémon', case=False)].quote.to_numpy()
```

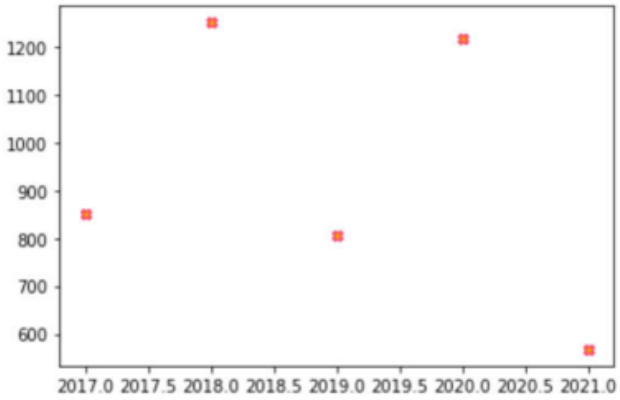
Out[13]:
```
array(['En mars 2017, un jeune Russe a été condamné à 3\xa0ans et
demi de colonie pénitentiaire pour avoir joué à Pokémon Go dans un
e église.'],
      dtype=object)
```

- Visualisation -

One plot of your choice, that you find relevant

In [25]:
```
plt.scatter(
    x=df.year.value_counts().index ,
    y=df.year.value_counts().values,
    c='#fcba03',
    marker = 'X',
    edgecolors = '#f51b6e'
)
```

Out[25]: <matplotlib.collections.PathCollection at 0x28fa45f7a90>



In [28]:
```
plt.bar(
    df.year.value_counts().index,
    df.year.value_counts().values,
    color = '#03b5fc',
    edgecolor = '#063547'
)
```

Out[28]: <BarContainer object of 5 artists>